Deconstructing Retrieval Abilities of Language Models

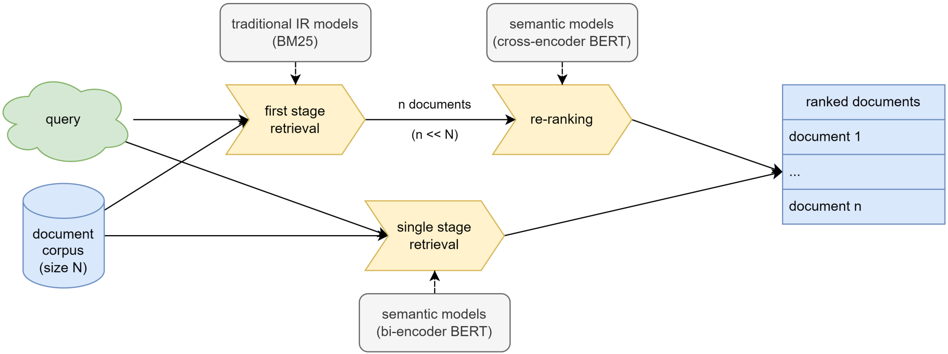
Hauke Tristan Hinrichs

Faculty of Electrical Engineering and Computer Science
Institute of Data Science
Department of Knowledge-based Systems
L3S Research Center

July 3, 2023

Motivation

- Information Retrieval (IR) decides which information is presented to us

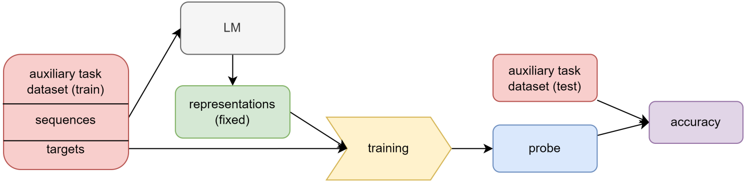


[1–3]

- Goal: shed light on inner workings of bi-encoder TCT-ColBERT [4, 5]

Motivation – Probing

- Probing: technique to *probe* for encoded information in the representations of language models (LMs) [6–8]

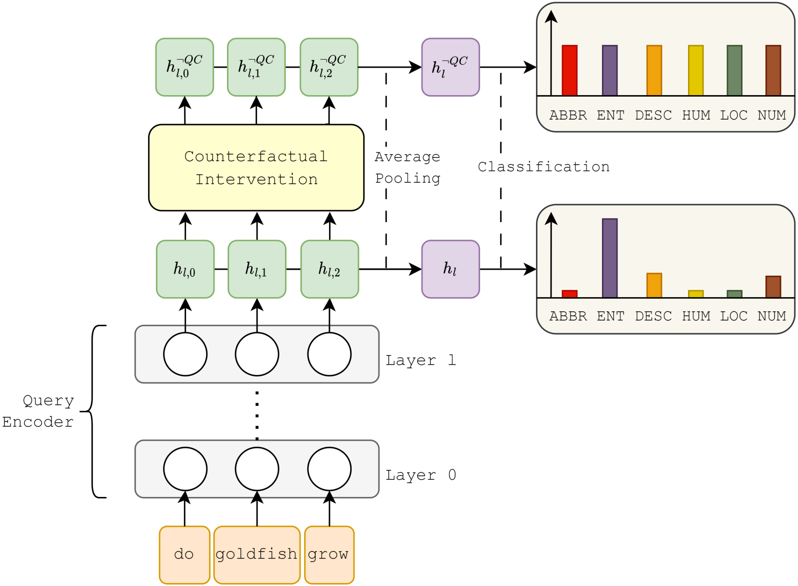


- Problem: encoding of information no proof for usage [9]
- *Causal Probing*: enabling causal explanations for model behavior by extending probing [10]

Research Questions

- **RQ1** Can we confirm the feasibility of *causally probing* our bi-encoder subject model in the context of retrieval?
- **RQ2** On which properties does our bi-encoder rely upon to solve the task of text retrieval?
- **RQ3** At which layers are important properties encoded?

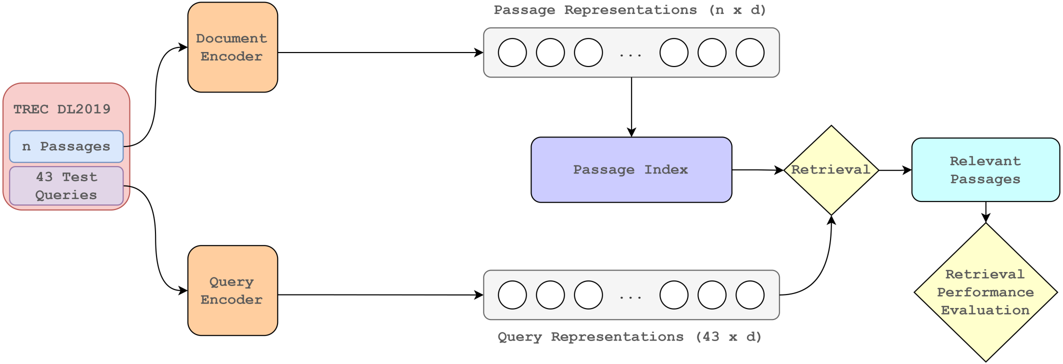
Approach – Causal Probing: Key Idea



Approach – Linear Adversarial Concept Erasure (R-LACE) [11]

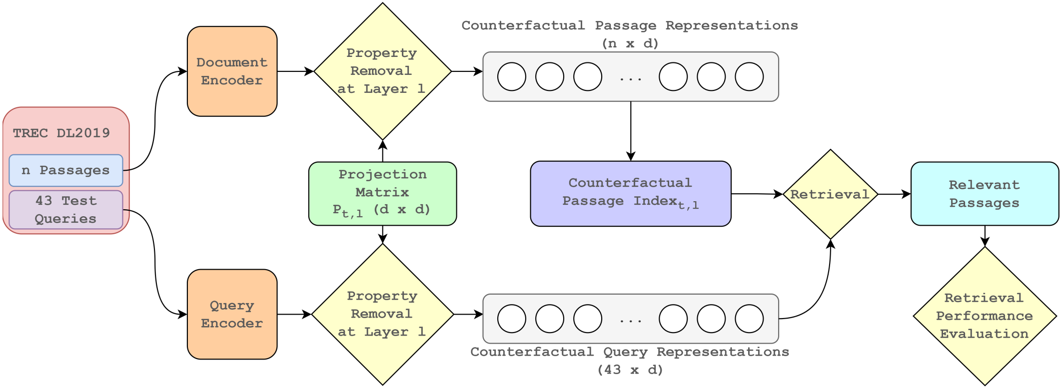
- Minimax game between two adversaries: linear predictor and a linear projection
- Goals:
  - Predictor unable to solve task in projected subspace
  - Minimal damage to unrelated information
- Input: Concept dataset, k (removed subspace rank)
- Output: Linear concept-removing projection

Approach – IR with Bi-Encoders



[12]

Approach – Causal Probing: Procedure



Approach – Investigated IR Properties

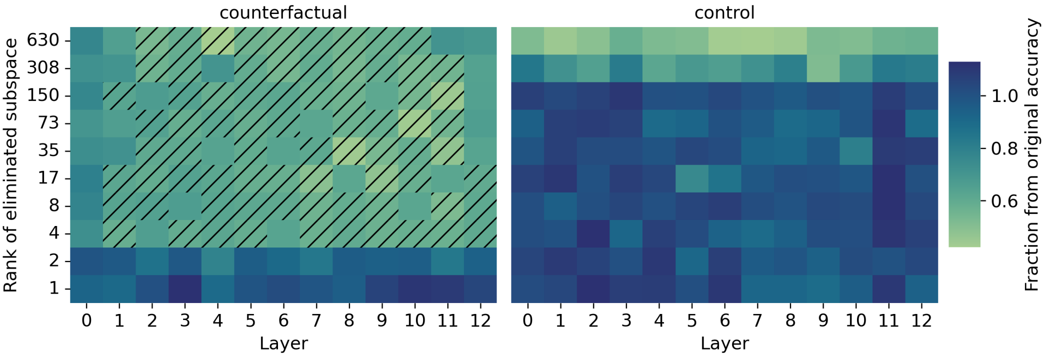
- **BM25**: exact term-matching [1] [2]
- **SEM**: Semantic similarity of query and document (cosine similarity between averaged GloVe-embeddings [13])
- **TI**: Term importance w.r.t. a query (RSJ weight) [14, 15]
- **NER**: Named-entity recognition
- **COREF**: Coreference resolution
- **QC**: Question classification

Approach – Feasibility Studies

1. Eliminating Subspaces of Increasing Ranks
   - Goals: Investigate influence of k; find the best k for each property
2. Probing as a Sanity Check
   - Goals: Confirm that properties are linearly encoded in the subject model's representations and R-LACE succesfully removes them

Results – Feasibility Study: Eliminating Subspaces of Increasing Ranks

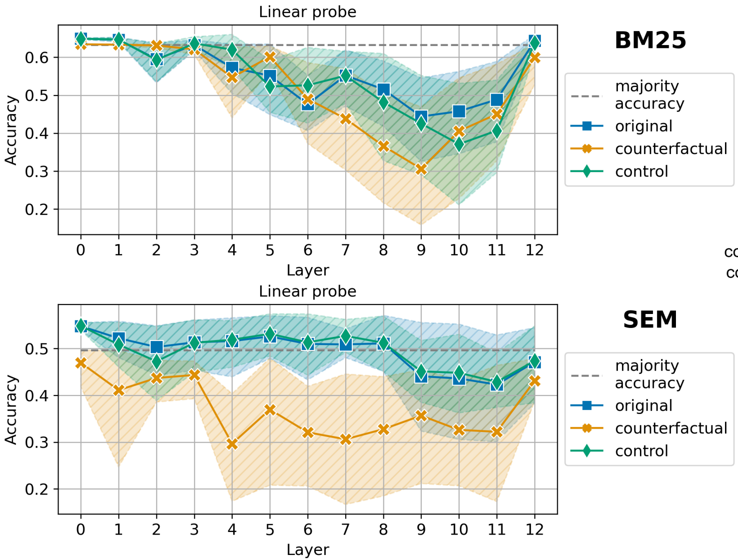
Depicted property: Question classification



Feasibility Study: Probing as a Sanity Check

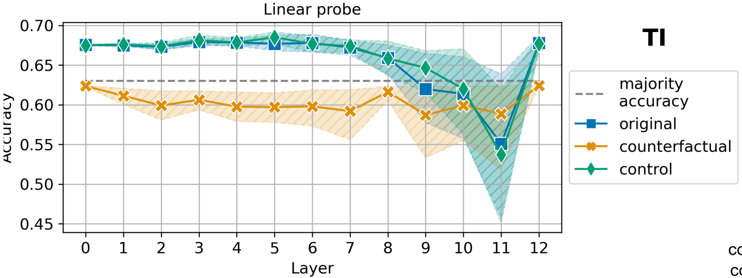
- Conventionally probe 3 kinds of representations for each property: original (fixed), counterfactual and control
- Sanity check considered passed when accuracies meet the following:
  1. original $>$ majority
  2. counterfactual $<$ original (preferably counterfactual $\leq$ majority)
  3. counterfactual $<$ control

Results – Feasibility Study: Probing as a Sanity Check (1/3)

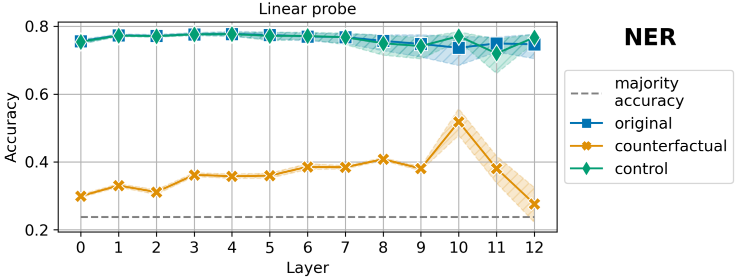


original > majority
counterfactual < original
counterfactual < control

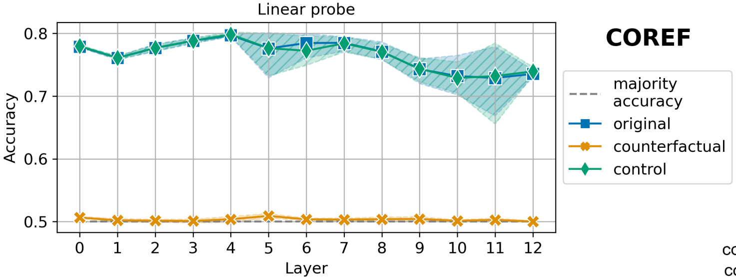
Results – Feasibility Study: Probing as a Sanity Check (2/3)



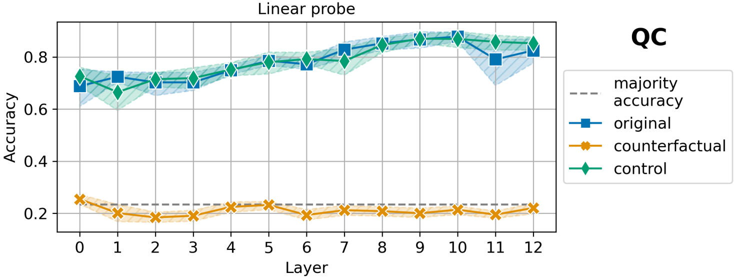
original > majority
counterfactual < original
counterfactual < control



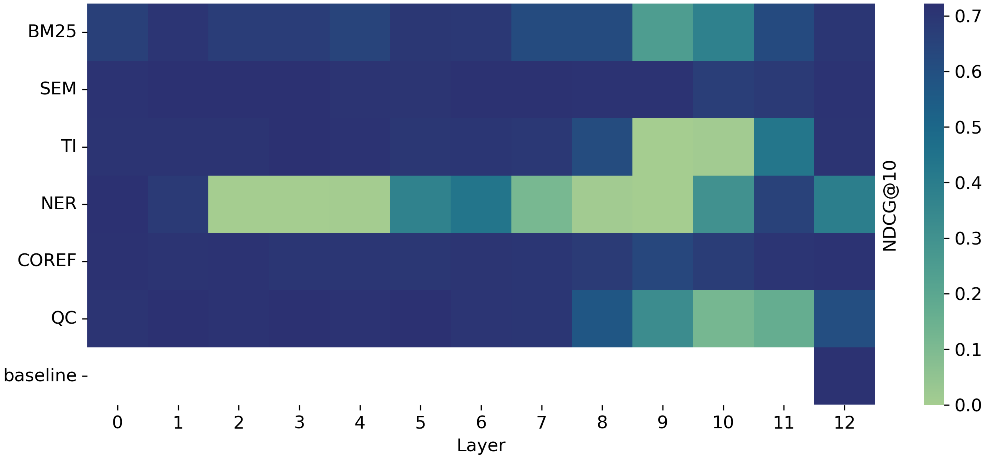
Results – Feasibility Study: Probing as a Sanity Check (3/3)



original > majority
counterfactual < original
counterfactual < control



Results – Causal Probing



Conclusion (1/2)

- **RQ1** Can we confirm the feasibility of *causally probing* our bi-encoder subject model in the context of retrieval?
  - Yes, for most of the properties. Limitations for BM25 and SEM.
- **RQ2** On which properties does our bi-encoder rely upon to solve the task of text retrieval?
  - Importance hierarchy: SEM, COREF $<$ BM25, QC $<$ TI, NER
- **RQ3** At which layers are important properties encoded?
  - Removal has larger impact at later layers, except for NER.

Conclusion (2/2)

- Limitations:
  - Only approximation of a property gets removed
  - Spurious correlations with a property
  - Only removal of linear information
- Future Work:
  - Additional properties
  - Investigate other bi-encoder architectures and training regimes
  - Use non-linear removal technique [16]
  - Use advancement of R-LACE: LEAst-squares Concept Erasure (LEACE) [17] (closed-form solution for complete linear concept erasure)

References I

[1] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, 01 2009. doi: 10.1561/1500000019. URL https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf.

[2] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Text Retrieval Conference*, 1994.

[3] Jimmy Lin, Rodrigo Frassetto Nogueira, and Andrew Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021. doi: 10.2200/S01123ED1V01Y202108HLT053. URL https://doi.org/10.2200/S01123ED1V01Y202108HLT053.

[4] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. Distilling dense representations for ranking using tightly-coupled teachers. *CoRR*, abs/2010.11386, 2020. URL https://arxiv.org/abs/2010.11386.

References II

[5] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In Anna Rogers, Iacer Calixto, Ivan Vulic, Naomi Saphra, Nora Kassner, Oana-Maria Camburu, Trapit Bansal, and Vered Shwartz, editors, *Proceedings of the 6th Workshop on Representation Learning for NLP, RepL4NLP@ACL-IJCNLP 2021, Online, August 6, 2021*, pages 163–173. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.repl4nlp-1.17. URL https://doi.org/10.18653/v1/2021.repl4nlp-1.17.

[6] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SJzSgnRcKX.

References III

[7] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1198. URL `https://aclanthology.org/P18-1198/`.

[8] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=BJh6Ztuxl`.

References IV

[9] Abhilasha Ravichander, Yonatan Belinkov, and Eduard H. Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL `https://doi.org/10.18653/v1/2021.eacl-main.295`.

[10] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL `https://doi.org/10.1162/tacl_a_00359`.

[11] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR, 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

[12] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820, 2020. URL https://arxiv.org/abs/2003.07820.

References VI

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014. doi: 10.3115/v1/d14-1162. URL `https://doi.org/10.3115/v1/d14-1162`.

[14] S. E. Robertson, S. Jones, and K. Spärck. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. doi: 10.1002/asi.4630270302. URL `http://dx.doi.org/10.1002/asi.4630270302`.

References VII

[15] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Match your words! A study of lexical matching in neural information retrieval. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 120–127. Springer, 2022. doi: 10.1007/978-3-030-99739-7_14. URL https://doi.org/10.1007/978-3-030-99739-7_14.

[16] Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6034–6055. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.405.

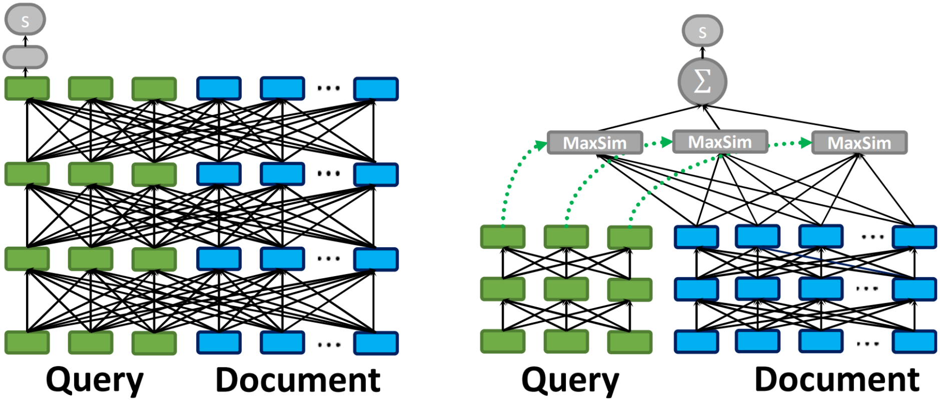
References VIII

[17] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: perfect linear concept erasure in closed form. *CoRR*, abs/2306.03819, 2023. doi: 10.48550/arXiv.2306.03819. URL https://doi.org/10.48550/arXiv.2306.03819.

[18] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Backup Slides

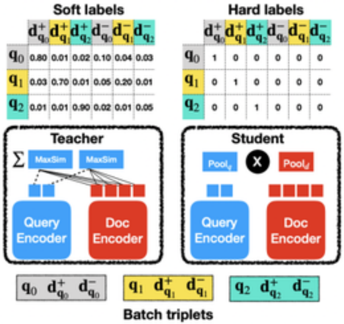
ColBERT [18]



(c) All-to-all Interaction
(e.g., BERT)

(d) Late Interaction
(i.e., the proposed ColBERT)

TCT-ColBERT [4, 5]



1. Teacher: ColBERT (BERT-based)

2. Student: BERT-based Bi-encoder with avg pooling

3. Student Training:

$$\mathcal{L} = -\sum_{i=1}^{|B|}\left\{\overbrace{\boxed{\gamma \cdot \log(P_S(d^+_{q_i}|q_i))}}^{\text{ranking loss}} + \underbrace{\boxed{(1-\gamma)\sum_{d' \in \mathcal{D}_B} KL(P_S(d'|q_i)||P_T(d'|q_i))}}_{\text{distillation loss}}\right\}$$

tight coupling: inference with the teacher while distillation, not beforehand

IR Properties – Examples

Task	Type	Level	Example
BM25	Regression	Sequence	**query**: most expensive hotels in new york city **passage**: The world's most expensive flight costs $38,000 — one way. Etihad Airways' new route connecting Mumbai and New York City... **target**: 22.063
SEM	Regression	Sequence	**query**: does insulin give you constipation **passage**: Summary: Constipation is found among people who take Insulin, especially... **target**: 0.132
AVG TI	Regression	Sequence	**query**: how long can ribs stay frozen **passage**: Raw pork chops can be safely frozen for up to six months... **target**: 2.763
TI	Regression	Token	**query**: where is hamvir's rest in skyrim **passage**: Hearthfire is the second DLC release for [**Skyrim**] behind the extremely successful... **target**: 10.336
NER	Classification	Token	**passage**: If you want to meet halfway between [**Los Angeles**], CA and Stockton, CA or just... **target**: Geopolitical entity (GPE)
COREF	Classification	Token	**passage**: [**Aluminum chloride**] is a chemical compound that has several uses, including as a treatment for excessive sweating and in antiperspirants. [**It**] is used... **target**: True
QC	Classification	Sequence	**query**: What is the full form of .com? **target**: Abbreviation (ABBR)

Linear Probe

- Binary or multinomial logistic regression model, depending on the task
- optimization goal (multinomial):

$$\min_{w,b} -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \frac{\exp(x_i w^{(k)} + b^{(k)})}{\sum_{j=1}^{K} \exp(x_i w^{(j)} + b^{(j)})} \tag{1}$$

NDCG

- main metric in TREC DL
- $$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}} \tag{2}$$
- $$\mathrm{DCG} = \sum_{i=1}^{|\mathcal{C}|} \frac{y_i}{\log_2(i+1)} \tag{3}$$

Term Importance – RSJ formula [15]

$$RSJ(t, q, \mathcal{C}) = \log \frac{p(t|\mathcal{R})p(\neg t|\neg\mathcal{R})}{p(\neg t|\mathcal{R})p(t|\neg\mathcal{R})} \tag{4}$$

Causal Probing Results – Recall@1000